

Data and Code for: Silence to Solidarity: How Communication About a Minority Affects Discrimination

Author: Duncan Webb (Nova School of Business and Economics)

Citation: Webb, Duncan. “Silence to Solidarity: How Communication About a Minority Affects Discrimination.” *Journal of Political Economy* (forthcoming).

Overview

This replication package constructs the analysis dataset from the raw survey data collected and undertakes the analysis. The replication package consists of a total of 33 scripts (including custom packages and functions used) in R and Python. The output generated by each file and the corresponding table in the main text is specified below.

Data Availability and Provenance Statements

Data Source: All raw data in this replication package was generated by the field experiment described in the paper. No external datasets were used.

The data was collected by the author through survey methods, with financial support from JPAL’s Crime and Violence Initiative, EUR-PgSE, the Weiss Fund, and CEPREMAP.

All of the raw data is provided for publication and included in this replication package post-anonymization. A detailed description of all the data files can be found in `data/Data Dictionary - Silence to Solidarity.xlsx`.

The analysis uses LLM generated data - the data used is provided as ‘raw’ data and code used to generate it is also included in the replication package.

All of the questionnaires used in the survey are also provided in the replication package.

Statement about Rights

- I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.
- The author will provide reasonable assistance to requests for clarification and replication.

Summary of Availability

- All of the non-anonymized raw data has been provided for data editor
- All data are publicly available post-anonymization of the raw data
- Data and code will be preserved for at least 5 years following journal policy

Pre-Registration

<https://www.socialscisceregistry.org/trials/10953>

Ethics

IRB approval documents are included in a separate folder in the replication package

Details on each Data Source

The table below documents the key raw data sources for the paper. A more detailed data dictionary of all the files can also be found in `data/Data Dictionary - Silence to Solidarity.xlsx`. A codebook with details on the variables can be found in `data/CODEBOOK.txt`.

The data dictionary includes a “Data Type” column that distinguishes between:

- **Raw:** Data collected from surveys or external sources (not generated by the analysis code)
- **Intermediary:** Data files generated by the analysis code (created from raw data through processing)
- **Output:** Final analysis outputs (tables, figures, etc.)
- **Reference:** Documentation files (questionnaires, etc.)

Data File	Location	Data Type	Provided for Publication	Description
GRI Main v2_wide.csv	data/confidential-not-for-publication/raw/	Raw	No	main survey results for phase 1
GRI Phase 2_wide.csv	data/confidential-not-for-publication/raw/	Raw	No	main survey results for phase 2
GRI Follow Up Delivery v3_WIDE.csv	data/confidential-not-for-publication/raw/	Raw	No	follow up survey for phase 1
GRI Follow Up Delivery - Phase 2_WIDE.csv	data/confidential-not-for-publication/raw/	Raw	No	follow up survey for phase 2
Trans Pilot_WIDE.csv	data/confidential-not-for-publication/raw/	Raw	No	supplementary data from pilot
Trans_recog_wide_anon.csv	data/cleaned/	Raw	Yes	anonimised supplementary data from pilot
randomisation_details.csv	data/raw/	Raw	Yes	contains details of randomisation for phase 1
randomisation_details_phase2.csv	data/raw/	Raw	Yes	contains details of randomisation for phase 2
corrections.xlsx	data/raw/	Raw	Yes	contains corrections of IDs
data_cleaned.RData	data/cleaned/	Intermediary	Yes	merged and anonimised dataset of the above files
data_cleaned.dta	data/cleaned/	Intermediary	Yes	merged and anonimised dataset of the above files
main_questionnaire_v3.xlsx	questionnaire/	Reference	Yes	questionnaire with data labels for phase 1
main_questionnaire_v4.xlsx	questionnaire/	Reference	Yes	questionnaire with data labels for phase 2
harmonise_reasons.csv	data/raw/	Raw	Yes	list of harmonised reasons for choice of delivery
harmonise_reasons_collapse.csv	data/raw/	Raw	Yes	list of harmonised reasons for choice of delivery
other_data_inputs.xlsx	data/raw/	Raw	Yes	Labels and socially desirable answers for the Crowne-Marlow social desirability score
lasso_control.RData	data/cleaned/	Intermediary	Yes	LASSO controls selected in the main specification
photo_rating_survey.sav	data/confidential-not-for-publication/raw/	Raw	No	dataset with responses to photo rating survey
photo_rating_survey_anon.sav	data/photo_rating_survey/	Raw	Yes	anonimised dataset with responses to photo rating survey
photo_IDs.xlsx	data/photo_rating_survey/	Raw	Yes	IDs for photo rating survey
photo_rating_survey_anon.csv	data/photo_rating_survey/	Intermediary	Yes	anonimised dataset with cleaned and merged responses to photo rating survey
transcripts_Madhavan.xlsx	data/confidential-not-for-publication/raw/transcripts/	Raw	No	raw discussion transcripts in Tamil
transcripts_Prakash.xlsx	data/confidential-not-for-publication/raw/transcripts/	Raw	No	raw discussion transcripts in Tamil
transcripts_Sathya.xlsx	data/confidential-not-for-publication/raw/transcripts/	Raw	No	raw discussion transcripts in Tamil
transcripts_IDs.xlsx	data/confidential-not-for-publication/raw/transcripts/	Raw	No	raw discussion transcripts in Tamil
translation_1.txt to translation_283.txt	data/confidential-not-for-publication/raw/transcripts_translations/	Raw	No	translated individual transcripts in English
transcripts_for_translation.csv	data/confidential-not-for-publication/cleaned/	Raw	No	intermediate file for translation purposes
transcripts_clean.RData	data/cleaned/	Intermediary	Yes	merged, cleaned, and translated transcripts with English and Tamil text
transcripts_clean.csv	data/cleaned/	Intermediary	Yes	merged, cleaned, and translated transcripts with English and Tamil text
completed_rows.csv	data/morality_analysis/	Intermediary	Yes	LLM generated dataset on analysis of moral topics
expanded_dataset.csv	data/morality_analysis/	Intermediary	Yes	LLM expansion of dataset on repetition of random moral topics
final_analysis.csv	data/morality_analysis/	Intermediary	Yes	LLM analysis of dataset on moral topics with ratings

Computational requirements

Software Requirements

- The replication package contains one or more programs to install all dependencies and set up the necessary directory structure.

R 4.5.0

- The program `Ob_functions.R` lists all packages used with the installation code if necessary
- Packages used - `tidyverse`, `qual`, `dups`, `trackr`, `readxl`, `knitr`, `modelsummary`, `rstatix`, `ggsignif`, `ggpubr`, `stringdist`, `lubridate`, `fixest`, `corrplot`, `xtable`, `kableExtra`, `ggpattern`, `tictoc`, `janitor`, `hdm`, `randomizr`, `ri2`, `ggplot2`, `ggmap`, `dplyr`, `haven`, `Hmisc`, `magrittr`, `ggh4x`, `tidytext`, `wordcloud`, `igraph`, `topicmodels`, `treemap`, `RColorBrewer`, `Rtsne`, `cluster`, `lsa`, `Matrix`, `httr`, `jsonlite`, `openai`, `psych`, `stringr`, `sandwich`, `lmtest`, `margins`, `ICC`
- The necessary package versions utilized in the scripts are stored in the `code/r-requirements.txt` file - please ensure the correct versions used

Python 3.14.0

- Packages used - `numpy`, `pandas`, `openai`, `cosine_similarity`, `pickle`

Brew

- Saving figures use a custom function which necessitates xquartz to be installed on your computer
- You can do so from their website or by running the following code from the command line/terminal: `brew install --cask xquartz`

Memory, Runtime, Storage Requirements

Summary

Approximate time needed to reproduce the analyses on a standard 2025 desktop machine:

- 2-8 hours if running without LLM code
- 8-24 hours if running with LLM code

Approximate storage space needed:

- 250 MB - 2 GB

Approximate memory (RAM) needed:

- 70 GB total (can use virtual memory/swap if physical RAM is insufficient)
- On systems with limited RAM, R can be configured to use disk-based memory by increasing swap space or using memory-mapped files

Details

The code was last run on a 10-core Apple M4 device with MacOS version 15.6.1 with 100GB of free space and 64GB of RAM (with additional swap space enabled).

Description of programs/code

- Master File - `0_master.R` will run all of the code scripts
- File `1_import.R` is provided for the data editor only as it utilizes non-anonymized data which is not shared for publication. The master file code will run
- File `17_import_transcripts.R` is provided for the data editor only as it utilizes non-translated data which is not shared for publication

Instructions to Replicators

- Edit `code/0_master.R` to adjust the working directory
- Custom packages and functions used are included in the script files `0a*` to `0d*`
- Set `rerun_llm <- TRUE` in the master file to redo analysis of LLM based analysis. Set to `FALSE` as default
- API keys
 - Parts of the code requires API keys for Google Maps and Open AI, specified below
 - * Google Maps:
 - File: `10_results_descriptives.R`, Line: 503
 - * Open AI:
 - File: `18_generate_embeddings.py`, Line: 7
 - File: `22_llm_rate_transcripts.R`, Line: 1
 - File: `24_ml_hypothesis_generation.R`, Line: 1
 - File: `25_ml_hypothesis_process.R`, Line: 17
 - Details of where to insert API keys are also stored in `code/API.txt`
- Download the data files referenced in the data dictionary. Each should be stored in the prepared & specified subdirectories of `data/`, in the format that you download them in. Do not unzip. No further action is needed on the replicator's part.
- Run `code/0_master.R` to run all steps in sequence.

Details of code files

- `0a_dups.R`, `0a_qval.R`, `0a_trackr.R`: Loads custom packages with the correct version

- `0b_functions.R`: Loads and installs (if necessary) all packages required in the code files. Contains custom functions used
 - Please ensure correct package version installation in `code/r-requirements.txt`
- `0c_lasso_controls.R`: Contains functions used in selecting lasso controls methods
- `0d_ri_functions.R`: Contains functions used to conduct randomization inference and calculate randomization inference p-values
- `1_import.R`: Loads raw datasets and merges them across phases, anonymizes and cleans datasets. **Only for data editor**
- `2_prep_dfs.R`: Cleans datasets and preps them for analysis
- `3_results_prep.R`: Sets up environment variables that are applicable to multiple scripts
- `4_result_main.R`: Main script for analysis results
- `5_results_videos.R`: Script for analysis of results related to videos shown regarding legality of discrimination
- `6_follow_up.R`: Script for analysis of follow up data & persistence of effects
- `7_results_mechs.R`: Main script for analysis of mechanisms for explaining results
- `8_results_appendix.R`: Script for results showed in the appendix.
 - Note - this script may prompt you to update the `margineffects` package, the code works irrespective of the option you choose. If you choose to update, there will be a prompt for selection where you should enter the CRAN mirror for your geographical region. If you choose to update it, the package may need to be downgraded to the correct version as per `r-requirements.txt` before running the code file again.
- `9_results_mechs_discussion.R`: Script for mechanism analysis and results
- `10_results_descriptives.R`: Script for creating descriptive tables and figures
- `11_balance.R`: Script for creating balance tables across phases and treatment
- `12_protocol_fidelity.R`: Script for checking robustness to issues with fidelity of the experimental protocol
- `13_weighting.R`: Script for creating sampling weights table
- `14_export_lasso_controls.R`: Script for lasso controls table
- `15_discussion_dynamics.R`: Script for analysis of discussion dynamics
- `16_photo_rating_survey.R`: Script for analysis of photo rating survey responses
- `17_import_transcripts.R`: Script for importing, cleaning and preparing transcript files. Generates intermediate files for translation as well. **Only for data editor**
- `18_generate_embeddings.py`: Script for analyzing transcripts and generating embedded quotes for further analysis
- `19_analyse_transcripts.R`: Script for analysis of transcripts
- `20_analyse_embeddings.R`: Script for analysis of embedded quotes in transcripts
- `21_het_by_posneg_discussion.R`: Script for analyzing heterogeneity by type of discussion
- `22_llm_rate_transcripts.R`: Script for using LLM to rate morality of transcripts
- `23_analyse_transcript_morality.R`: Script for analyzing morality rating generated by LLM
- `24_ml_hypothesis_generation.R`: Script to generate hypotheses about differences between transcripts, using LLMs
- `25_ml_hypothesis_process.R`: Script to test hypotheses about differences between transcripts, using LLMs
- `26_discussion_debrief.R`: Script for analyzing results of post-discussion debriefs
- `27_intensity_pref.R`: Script for individual level willingness to pay to choose transgender worker

List of tables and programs

Below is a list of all the tables and figures from the text along with the stored location in the replication package and the code file that generates it. The first part shows the locations and file names for the main text tables and figures while the second part of the table contains the same for the Appendix.

Main Text

Fig/Table #	Program	Line	Output file
Table 1	<code>4_result_main.R</code>	199	<code>outputs/tables/main_table.tex</code>
Table 2	<code>5_result_videos.R</code>	119	<code>outputs/tables/videos.tex</code>
Table 3	<code>6_follow_up.R</code>	67	<code>outputs/tables/follow_up.tex</code>
Table 4	<code>4_result_main.R</code>	656	<code>outputs/tables/mechs_combine_r1_r2.tex</code>
Table 5	<code>9_results_mechs_discussion.R</code>	858	<code>outputs/tables/effect_of_announce_sob_discrim.tex</code>
Table 6	<code>23_analyse_transcript_morality.R</code>	178	<code>outputs/tables/models_morality.tex</code>
Figure 4	<code>4_result_main.R</code>	124	<code>output/figs/main_graph_point.pdf</code>
Figure 5	<code>21_het_by_posneg_discussion.R</code>	481	<code>outputs/figs/het_by_posneg_discussion.pdf</code>

Appendix

Output File	Program	Line	Tex File Reference
outputs/stats/spoke_pro_trans_mean.tex	26_discussion_debrief.R	471	5357
outputs/stats/spoke_pro_trans_nobs.tex	26_discussion_debrief.R	251	5357
outputs/stats/prop_choosing_only_saw.tex	10_results_descriptives.R	585	5885
outputs/stats/prop_choosing_only_spoke.tex	10_results_descriptives.R	591	5885
outputs/stats/prop_listener_silent.tex	10_results_descriptives.R	611	5885
outputs/stats/prop_others_could_hear_answers_rev.tex	10_results_descriptives.R	604	5887
outputs/stats/salience_trans.tex	10_results_descriptives.R	263	6399
outputs/stats/salience_non_trans.tex	10_results_descriptives.R	265	6399
outputs/stats/list_exp_coeff_list_a.tex	10_results_descriptives.R	355	6405
outputs/stats/list_exp_coeff_list_b.tex	10_results_descriptives.R	360	6405
outputs/stats/list_exp_p_diff.tex	10_results_descriptives.R	373	6405
outputs/stats/prop_an_check1.tex	10_results_descriptives.R	649	6438
outputs/stats/prop_an_check2.tex	10_results_descriptives.R	654	6438
outputs/stats/ls_corr_basic.tex	10_results_descriptives.R	186	6451
outputs/stats/ls_corr_basic_p.tex	10_results_descriptives.R	192	6451
outputs/stats/ls_corr_fes.tex	10_results_descriptives.R	211	6451
outputs/stats/ls_corr_fes_p.tex	10_results_descriptives.R	222	6451
outputs/stats/item_sensitivity.tex	4_results_main.R	234	6705
outputs/stats/statistical_discrim.tex	8_results_appendix.R	219	6713
outputs/stats/diff_mf.tex	8_results_appendix.R	2523	6717
outputs/stats/diff_mf_p.tex	8_results_appendix.R	2524	6717
outputs/stats/diff_mf_disc_p.tex	8_results_appendix.R	2527	6717
outputs/stats/eff_female.tex	8_results_appendix.R	2643	6717
outputs/stats/eff_female_p.tex	8_results_appendix.R	2644	6717
outputs/stats/eff_female_male_only.tex	8_results_appendix.R	2655	6717
outputs/stats/eff_female_male_only_p.tex	8_results_appendix.R	2656	6717
outputs/stats/p_vals_video_placebo_discussion.tex	5_results_videos.R	651	6833
outputs/stats/effect_belief_vid.tex	5_results_videos.R	380	6838
outputs/stats/effect_belief_vid_perc.tex	5_results_videos.R	382	6838
outputs/stats/coeff_amount_discussed_listener_trans.tex	9_results_mechs_discussion.R	1769	6896
outputs/stats/amount_discussed_listener_trans.tex	9_results_mechs_discussion.R	1770	6896
outputs/stats/pval_disagreement.tex	19_analyse_transcripts.R	32	6901
outputs/stats/anon_trans_penalty.tex	7_results_mechs.R	1556	6936
outputs/stats/anon_effect.tex	7_results_mechs.R	1559	6938
outputs/stats/anon_effect_p.tex	7_results_mechs.R	1562	6938
outputs/stats/effect_duration_discussion.tex	8_results_appendix.R	2301	7111
outputs/stats/effect_duration_discussion_perc.tex	8_results_appendix.R	2302	7111
outputs/stats/p_val_dominated_mf_r2.tex	8_results_appendix.R	1933	7111
outputs/stats/p_val_corr_duration_choice.tex	8_results_appendix.R	2360	7111
outputs/stats/purpose_2_mean.tex	7_results_mechs.R	1412	7277
outputs/stats/p_val_mem_check_diff_trans.tex	8_results_appendix.R	1570	7298
outputs/stats/high_stakes_p_val.tex	8_results_appendix.R	1064	7312
outputs/stats/f_purpose_0.tex	10_results_descriptives.R	96	7403
outputs/stats/f_purpose_0_p.tex	10_results_descriptives.R	97	7403
outputs/stats/f_purpose_2.tex	10_results_descriptives.R	99	7403
outputs/stats/f_purpose_2_p.tex	10_results_descriptives.R	100	7403
outputs/stats/n_high_stakes_sample.tex	8_results_appendix.R	1069	7450
outputs/stats/n_low_stakes.tex	8_results_appendix.R	1079	7450
outputs/stats/n_high_stakes.tex	8_results_appendix.R	1075	7450
outputs/stats/discussion_length_words_lead_trans.tex	23_analyse_transcript_morality.R	218	7466
outputs/stats/discussion_ratio_lead_trans.tex	23_analyse_transcript_morality.R	223	7466
outputs/stats/discussion_length_words_lead_non_trans.tex	23_analyse_transcript_morality.R	219	7466
outputs/stats/discussion_ratio_lead_non_trans.tex	23_analyse_transcript_morality.R	224	7466